DOCUMENT RESUME

ED 203 714 FL 012 402

AUTHOR Clark, John L. D.
TITLE Language Testing and Assessment.
SPONS AGENCY National Inst. of Education (ED), Washington, D.C. National Council on Education Research.
PUB DATE 29 Apr 81
CONTRACT NIE-P-81-0056
NOTE 29p.: Paper presented at a meeting of the National Council on Education Research (Washington, DC, April 29, 1981).

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Academic Achievement; Agency Role; Educational Assessment; Elementary Secondary Education; Information Dissemination; *Language Proficiency; *Language Tests; Postsecondary Education; *Second Language Instruction; Testing
IDENTIFIERS American Council on the Teaching of Foreign Langs

ABSTRACT

This is one of ten papers commissioned to investigate research needs and to stimulate planning activities. The paper assumes that the primary current goal of foreign/second language instruction in the United States is to produce citizens who have a usable command of one or more languages other than their native language. The implications of this assumption are reflected in the need to develop instruments and procedures which will: (1) determine student attainment of pragmatically useful levels of language proficiency, (2) monitor student progress toward proficiency-based goals and provide appropriate instructional feedback, and (3) provide information concerning the particular instructional procedures and curricula best suited to individual language learners. Of these three areas, it is urged that the National Institute of Education (NIE) pay immediate attention to the measurement of language proficiency outcomes. With respect to course-of-instruction achievement measurement, attention is drawn to potential NIE support of a "realistic learning outcomes" project recently proposed by American Council on the Teaching of Foreign Languages. The development of a uniform, general purpose biographical/language history questionnaire by NIE is also advocated, as well as the establishment of a test information clearinghouse. (Author/AMH)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED203714

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

Language Testing and Assessment

John L. D. Clark

Educational Testing Service

It is not possible to discuss the present "state of the art" in foreign/second language assessment or to outline needed research and development activities in this area without first identifying what, for purposes of this paper, will be taken as the major intended outcome, on a broad policy basis, of foreign/second language instruction within the United States within the present decade and presumably well beyond: to produce American citizens who have a "real-life usable" command of one or more languages other than their native language. This major purpose of language instruction is consistent with the general tenor and specific recommendations of the President's Commission on Foreign Language and International Studies (1979), as well as with the expressed positions of major language organizations as represented, for example, in the MLA-ACLS Task Force Report on Language Study for the 1980's (Brod, 1980) and the recommendations arising from the National Conference on Professional Priorities at the 1980 annual meeting of the American Council on the Teaching of Foreign Languages (report in press). Beyond these organizational endorsements, lay public, student, and business-area expectations for the outcomes of foreign/second language instruction also provide strong support of proficiency-based language learning goals.

Although the following discussions will take a basic "proficiency" orientation as a given, it is important to emphasize that other potential goals of language instruction--expressed in terms of personal development, increased cultural/humanistic awareness, and other cognitive or affective outcomes--are

FL012402

certainly not incompatible with the development of language proficiency per se and indeed may be expected to interact with the latter on a synergistic basis.

If the acquisition of "real-life usable" language proficiency can be considered a primary goal of present-day language instruction in the United States, it becomes immediately possible to identify a number of significant implications of this orientation for language assessment activities in the service of this goal.

First, means are needed to determine student attainment of "real-life usable" proficiency as a result of the instructional process. Information obtained from this end-point assessment is of importance to individual students as confirmation of the practical success of their learning efforts; to schools and colleges as a quality control measure for their language teaching programs; and, on a nationwide basis, to policy makers in areas for which the national yield of language-competent students is an important analysis and planning consideration.

Second, in the course of the instruction itself, testing procedures capable of monitoring, at fairly short and repeated intervals, the direction and adequacy of progress toward the eventual proficiency goals are called for, both for informational and motivational feedback to the students and for guidance of the teacher in planning subsequent instructional activities.

Third (and closely related to the second) is the development and use of assessment procedures providing what might be called "starting-off" data--that is, information concerning the student's background, prior language experiences, personal interests in language learning, cognitive style, and other variables that would, ideally, be taken into account both in placing the student

appropriately within the instructional stream and subsequently arranging for him or her to follow those particular instructional options which would be expected to maximize acquisition of the intended proficiencies.

The three major assessment approaches briefly outlined above may usefully be referred to as involving, respectively, outcome measurement, process measurement, and input measurement activities, and these labels will be adopted in the following discussion. For each assessment category, the general "state of the art" will be described--from the viewpoint of proficiency-oriented language instruction--and areas of insufficient research and/or needed development work identified.

INPUT MEASUREMENT

As previously suggested, the basic purpose of prognostic "input" measurement is to obtain, at a given point in time, student-related information that will be of value in making appropriate pedagogical choices from among various alternative courses of action. With regard to the nature and role of student input variables in language acquisition, the earliest and as of the present time most detailed work has been in the area of "language aptitude" measurement. Primary efforts in this area were made by John B. Carroll, who in the mid-1950's worked with the U.S. Army Language School and other government language training programs to design and validate a series of measures that were subsequently found to predict, with a fairly high level of precision, student learning success in intensive foreign language courses of the types offered in these institutions (Carroll, 1962). Further studies of the predictive efficiency of the Carroll measures in regular academic language courses at both secondary and college levels also showed positive, though in general somewhat lower, correlations with course grades and other measures of language learning success.

A published version of the Carroll aptitude battery has been available since the late 1950's as the Modern Language Aptitude Test (Carroll and Sapon, 1959), and in a modified form intended for elementary school students (MLAT-Elementary) since 1967 (Carroll and Sapon, 1967). A third aptitude testing battery, generally similar to the MLAT, was developed at about the same time by Pimsleur and published as the Language Aptitude Battery (LAB) (Pimsleur, 1966).

Although language aptitude measurement as exemplified in--and given operational definition by--instruments of the Carroll and Pimsleur type continues to play a very useful role in specialized learning situations and as a control variable in methodological studies, several considerations have reduced the importance of this type of assessment for widespread use at the present time, for both general language-policy and psychometric reasons. With respect to language policy considerations, except for a relatively few language teaching situations in which a high level of student selectivity is necessary and appropriate (again, largely in specialized intensive language programs), present-day emphasis is on making appropriate foreign/second language learning experiences available to any students needing or desiring such instruction--in a very real sense regardless of level of language "aptitude" as measured by the conventional language aptitude batteries.

Consistent with this policy focus, and supported by an expanding body of relevant practical and research-based knowledge, is the developing view of language learning success as a function of a large number of other-than-aptitude variables that operate in the learning context. The work reported in Gardner and Lambert (1972), Gardner, Gliksman, and Smythe (1978), and elsewhere on the role of attitudinal factors, levels of motivation for language learning, perceived

purposes of language study, and other student-related variables as mediators of learning success points to a richness and complexity in "what the student brings" to the instructional situation that will require a considerably more wide-ranging assessment approach than can be provided by basic language aptitude measures per se.

Continued work in the prognostic assessment area--within an investigative framework which would undertake to represent as fully as possible the many different variables that are now clearly seen to be at issue--would be of considerable value from a basic research standpoint. However, a major impediment to making productive current or near-future use of refined prognostic information (if it were indeed available) is the lack of sufficient curricular flexibility to offer those specific types of language learning experiences that would be identified as optimal for a particular student or group of students. In view of the relatively few secondary school or university settings in which more than a single uniform language learning program would be an administratively or financially realistic possibility at the present or in the near-term future, the societal "payoff" of having highly sophisticated prognostic instruments available must await corresponding curricular options that may be some time away in their development.

OUTCOME MEASUREMENT

Of the three broad areas of assessment under discussion, effective measurement of the intended outcome goal of real-life usable proficiency on the part of foreign/second language learners plays the most salient and crucial role, since it constitutes both the evaluative benchmark against which language learning is measured and the basic focus toward which the instructional process must be continually and closely directed if this instruction is to achieve the

desired pedagogical result. With respect to their use as benchmark instruments, appropriately designed, administered, and interpreted proficiency tests can accomplish all of the following:

(1) provide important data concerning the overall yield of language instruction on a nationwide basis;

(2) (in suitably designed and implemented studies) serve as external, unbiased measures of the relative success or lack of success of different language learning curricula or methodologies;

(3) inform and motivate both students and teachers by providing meaningful information concerning the extent to which their learning/teaching efforts have resulted in the acquisition of pragmatically useful levels of language competence; and

(4) serve the very important (and currently lacking) role of a readily understandable lingua franca in communicating among the educational, business/commercial, and general public sectors concerning language competency needs and expectations "in the field" and the extent to which these are being met by the language teaching effort.

With regard to their use as guideposts for instruction, it is crucial that the criterion proficiency tests provide for highly valid measurement of the competencies in question. Otherwise, to the extent that the proficiency tests fail to relate closely to real-life language use, teaching approaches aimed at producing a high level of performance on these instruments run the risk of failing to inculcate the desired competencies or, at best, of doing so on a largely coincidental and inefficient basis. In much the same way, unless the criterion proficiency tests are highly valid, course-of-instruction process tests intended to monitor students' short-term progress toward the

proficiency goals will risk straying from the intended objectives. Faulty proficiency tests will also badly serve further research and development work in the "input" measurement area, in the sense and to the degree that the logical and statistical processes involved in the prognostic testing research will thus be aimed at maximizing the prediction of inappropriate criteria.

In view of the crucial position of the criterion proficiency measures within the total language teaching and assessment process, an important immediate consideration is the extent to which language proficiency tests developed to date may be considered valid for the purposes to be served. Within the United States, the "golden era" for the preparation of foreign/second language tests aimed at general proficiency measurement (i.e., measurement independent of a given text-book series or instructional program) was from the late 1950's to mid-1960's, coinciding with and for the most part funded through the National Defense Education Act. During this period were developed both the MLA Foreign Language Proficiency Tests for Teachers and Advanced Students (Starr, 1962)--covering the four skill areas of listening comprehension, reading comprehension, speaking, and writing in the five major languages of French, German, Italian, Russian, and Spanish--and the lower-level MLA-Cooperative Foreign Language Tests, also in the same four language skill areas and five languages, and intended for use at the high school and beginning college levels (Bryan, 1966; see also Clark, 1965 for a detailed technical review). A battery of generally similar tests in French, German, and Spanish was developed by Pimsleur and made available commercially in the mid-1960's (Pimsleur, 1967).

All three of these batteries displayed useful and innovative qualities by comparison to previously available instruments--most significantly, in providing for the direct testing of the productive skills of writing and speaking

in addition to the passive (and more easily testable) competencies of listening comprehension and reading. In addition, especially within the MLA Proficiency Test battery, serious efforts were made to provide realistic textual materials in both the listening comprehension and reading sections.

Although these efforts were laudatory and a significant step forward in the language assessment area generally, the resulting instruments--and the interpretive information which they provided--must be considered to have fallen short, in several respects, of the goal of providing appropriate and useful measures of "real-life" language performance. For example, with respect to the reception of spoken language, the listening comprehension section of all three test batteries consists of tape recorded single sentences, dialogues, and short passages spoken under quite artificial (controlled studio) conditions, with carefully enunciated speech, and do not closely parallel the actual conditions under which speech of various types would be received in the real-life settings involved. With respect to language production on the part of the examinee, the speaking tests in all three batteries make use of a tape recording/booklet administration format in which live conversational interaction and the communicatively significant real-time alterations in the speech content on the part of the interlocutors cannot be provided for. Important socio-linguistic components of real-life communication (e.g., conveyance of anger, doubt, humor, or other intended affect; selection of appropriate tone and register for the speech situation involved) are not addressed in the tests, nor are there any explicit attempts to assess the examinee's use of conversational management strategies (e.g., paraphrasing, requests for repetition or rephrasing), all of which have an important bearing on communicative success in real-life settings.

Beyond the questions of test content and testing procedure as such, the most serious drawback to the use of the MLA and Pimsleur tests and other similar instruments as effective measures of real-life language competence is the lack of detailed interpretive information which, in an accurate and straightforward manner, relates scores obtained on these instruments to the examinee's ability or lack of ability to carry out given communicative/receptive tasks in the language. Although the percentile norms originally provided with these test batteries did permit inter-examinee comparisons of the general "student X is more proficienct than student Y" type, there were not (nor are there presently available) any significant bodies of interpretive data characterizing the specific language-use tasks that examinees scoring at different levels could actually perform outside of the classroom setting.

It is appropriate at this point to describe one available and fairly widely used testing procedure which approaches somewhat more closely the testing format and manner of scoring and interpretation that are implied in the concept of "real-life referenced" proficiency tests: this is the so-called "absolute proficiency rating scale" and interviewing procedure developed by the U.S. Foreign Service Institute in the mid-1950's (Rice, 1959; Wilds, 1975; Sollenberger, 1978), and commonly referred to as the "FSI interview." This procedure incorporates a live conversational setting, thus providing for communicative interaction of the two interlocutors, and the structure of the interview is such that within a relatively short time span a fairly large number of language-use situations can be presented and evaluated. Scoring is based on a six-category scale ranging from no functional proficiency (level 0) up to the proficiency of an educated native speaker of the language (level 5).

Most significantly, unlike the MLA and other norm-referenced tests, the FSI scores are reported--and interpreted--by reference to short verbal descriptions of the kinds of language-use situations in which an examinee scoring at that level would be considered able to operate in an effective and appropriate manner. For example, the verbal description associated with "level 2" performance is as follows:

> Able to satisfy routine social demands and limited work requirements. Can handle with confidence but not with facility most social situations including introductions and casual conversations about current events, as well as work, family, and autobiographical information; can handle limited work requirements, needing help in handling any complications or difficulties; can get the gist of most conversations on non-technical subjects (i.e., topics which require no specialized knowledge) and has a speaking vocabulary sufficient to express himself simply with some circumlocutions; accent, though often quite faulty, is intelligible; can usually handle elementary constructions quite accurately, but does not have thorough or confident control of the grammar.

The FSI interview is by no means an ideal proficiency measuring instrument; for example, it does not lend itself well to the elicitation of interrogative expressions (cf. Lado, 1978), and tends to combine both linguistic (i.e., syntactical and morphological) and general communicative criteria within the individual proficiency level descriptions, as can be noted above. Nonetheless, the FSI interview and scoring scale represents one of the closest approximations of "real-life" proficiency assessment available at the present time.

Within the past several years, a number of organizations and individual researchers working within the modern language program of the Council of Europe (Trim, 1971, 1973; Van Ek, 1975) have developed what has come to be termed the "functional/notional" approach to curriculum design, in which learning objectives are based on a detailed specification of the language-use situations in which the learner will be expected to operate, the social functions to be performed (e.g., requesting, apologizing, questioning), and the specific "notions" to be

conveyed (e.g., requesting something to eat, apologizing for being late, etc.). An important merit of the functional/notional approach as it relates to language proficiency testing is the continued explicit emphasis which it places on real-life language use in specific communicative settings. By the same token, however, a potential drawback of the functional/notional concept is the virtually limitless number of possible language-use situations that could be specified in this manner and for which a separate, tailor-made assessment procedure would, at least in theory, be required. In order to provide any degree of practical guidance for proficiency test development, it would appear necessary to identify "meta-sets" of language elements common to a number of individual functions (and the associated communicative notions) in order to provide a more realistically manageable corpus of material from which to develop the proficiency test content.

In summary of this very brief overview of the proficiency testing situation as of the present time, it may be suggested that we are currently in a "holding pattern"--having developed, administered, and found somewhat wanting an earlier generation of test instruments, having discerned worthwhile avenues of further exploration in interview-based approaches to speaking testing and in functional/notional concepts, but not having gone appreciably beyond this point in conceptualizing and conducting basic research on (and much less, operationally developing) instruments and procedures that could come considerably closer to the proficiency measurement ideal.

In view of the crucial importance of adequate measures of language proficiency for real-life purposes--not only as criterion measures of instructional outcomes but also as necessary components of associated efforts in curriculum development, achievement testing, and prognostic measurement, it is

recommended that the NIE consider assuming a major role in working out the framework for and implementing a comprehensive research and development program of proficiency-based foreign/second language testing. The two major functions of such a program would be to (1) develop a variety of proficiency measurement instruments and associated procedures capable of providing scores that can be referenced to effectiveness of performance in actual language-use situations; and (2) following development of the tests themselves, provide a means for the accumulation and dissemination of feedback data relating test scores to empirically determined performance in a large variety of language-use contexts.

Although alternative approaches might be possible, the administrative framework within which the necessary project work could most effectively take place would probably involve the establishing of a "research and development center on language proficiency testing." Such a center would presumably be associated with a higher education institution, testing organization, or other established entity having suitable prior experience, staff expertise, and resources. Within such a structure, research-based development of prospective testing procedures, their initial tryout and modification into operational versions, and, importantly, the ongoing accumulation and dissemination of relevant interpretive information, could be carried out on a carefully planned and thoroughgoing basis, with natural and efficient linkages between the various project phases.

Although it is not possible to describe at this time the detailed nature and scope of the research/development activity that would be involved in the proficiency testing project, it is anticipated that this would include the following major phases:

(1) Assembly of an extensive battery of experimental test formats and assessment procedures. At the present stage of development of the language testing science, there is no single body of theory and supporting "hard" data that would point directly and unambiguously to the most valid and useful testing formats/procedures for language proficiency testing. The previously cited Council of Europe activities--together with recent efforts to develop frameworks of communicative competence by Canale and Swain (1979), B.J. Carroll (1980), and the "Common Yardstick" working group currently operating through ETS with Department of Education (International Studies branch) assistance--would be expected to provide some useful guidance concerning both the total scope of linguistic/sociolinguistic elements to be included in the tests and the general testing approaches to be employed. However, a good proportion of eclecticism would also be required in order to identify the greatest possible number of potentially relevant techniques and formats for inclusion in an extensive experimental battery of proficiency-oriented tests to be submitted to large-scale trial administration on common groups of examinees.

(2) Concurrent development of external validation measures. A basic theoretical question in proficiency test development (and one which also has substantial practical ramifications) is the nature of the criterion or criteria against which the proficiency tests themselves are to be validated. Except for the long-term surreptitious observation of examinees engaged in language-use tasks in the course of their daily activities, there is no way to formally assess "real-life performance" without departing, at least to some extent, from the actual language-use situations in question. For example, even in the FSI-type interview, the "natural" conversational context is still quite obviously

a testing situation, with affective and psychological influences operating on the examinee in ways that are different from those involved in the "genuine" communication situations that the interview is intended to reflect.

Rather than attempting to closely simulate real-world situations and to validate the proficiency tests against performance in these situations, a more productive strategy for obtaining validity-related information would be to make use of the two potentially very powerful tools of proficiency self-report (i.e., the examinee asked to appraise his or her language proficiency against specified performance standards) and second-party reports by individuals in the field who have had the occasion to observe and analyze the examinee's language performance at close range and on an extended basis in the real-life settings in question (for example, employment supervisors in second-language-using job settings).

Recent experimental evidence from other sources (Clark and Swinton, 1979; Barrows, Klein, and Clark [in press]) suggests that properly developed self- and second-party appraisal instruments can provide quite sensitive and reliable information on language performance which could be used, in conjunction with face- and content-validity considerations, in the construct validation of instruments developed under the proposed testing project.

(3) Large-scale validation administrations. This stage would involve administration of the experimental testing battery, as well as the criterion validation instruments, to a large number of examinees drawn from a variety of potential examinee groups (e.g., students in regular school language programs, language teachers, residents and tourists abroad, incumbents in a number of different language-related jobs, and others) in order to determine--on the basis of both the statistical results of the administrations and the observed practical

advantages and disadvantages of the various testing approaches--subsets of formats and procedures that most accurately and effectively predict the self-rating and second-party criteria, as well as interrelate among themselves on a conceptually meaningful basis. Two different subset batteries might be developed through these administrations: comprehensive "long form" tests to be used in situations where the most highly precise and reliable information is required (and where the correspondingly greater administration time and expense is justified), and "short form" tests for use where a useful degree of informational feedback is needed but where the realities of the testing situation do not permit the more comprehensive approach.

(4) Provision of ongoing interpretive information. Once the proficiency tests were made available for operational use, it would be very important for a test information clearinghouse or other suitable facility to be established, through which use-related data from the field would be collected and disseminated to other test-using groups. For example, summer-abroad or year-abroad programs using the tests as selection devices could report to the clearinghouse (with appropriate provisions for maintaining the confidentiality of individual examinees) student scores on the tests as related to in-country language performance as judged by the program staff. Analogous types of information could of course be provided and reported in a great number of other situations, including both academic and employment contexts. The importance of continuing to provide for use-related data gathering and reporting cannot be overemphasized, since it is precisely in order to obtain this type of information that the proficiency tests would have been initially developed. Such information, broadly disseminated and widely used for socially important purposes, would represent the fundamental "return-on-investment" of the entire test development process.

The implementation of a comprehensive proficiency testing program along the general lines suggested above (see Clark, 1980 for more detailed discussion) would be very much in keeping with the recommendations of the President's Commission on Foreign Language and International Studies for the establishment of a "National Criteria and Assessment Program" (Ch. 1, Sect. V), whose purposes, among others, would be to "assess the proficiency of both students and teachers in existing as well as new or experimental foreign language programs;" "monitor application of language skills by students on completion of their training;" and "determine the language skills required to function in a variety of contexts." To accomplish each of these purposes, as well as to provide the numerous other educational and societal benefits that would accrue from having in place (and widely utilizing) effective measures of the important end products of foreign/second-language instruction, a major NIE initiative in the area of language proficiency measurement could provide the strong and extremely timely impetus needed to attain such a goal.

PROCESS MEASUREMENT

As indicated earlier, the basic purpose of process measurement is to provide useful feedback information, at relatively short time intervals, concerning the extent to which students are acquiring each of a number of constituent elements of instruction which, in total and on a properly sequenced basis, would be expected to lead them up to the desired outcome performances measured by the real-life referenced proficiency tests described in the preceding section.

Of the three types of assessment under discussion, process testing may be considered to have the most extensive background of study and development. The important pioneering effort in this area was the well-known language testing volume of Robert Lado (1961), in which a contrastive analysis approach was used

as the framework for elaborating series of highly focused question types aimed at verifying the student's acquisition or lack of acquisition of specified elements of phonology, morphology, and syntax in both productive and receptive modes. This publication was followed over a period of years by additional works, aimed mainly at process testing, by Valette (1967), Harris (1969), Clark (1972), Valette (1977) and, most recently, Oller (1979) and Cohen (1980). The theoretical approaches and recommendations in these volumes and related journal articles and other publications over this period have undergone a certain amount of evolutionary change. Branching out from the highly particularistic "individual language element" model followed for the most part by Lado, later authors in the field have introduced considerations of "integrative" testing based on larger and more natural units of language material and have urged attention to the communicative effectiveness of the student's performance as well as to questions of linguistic accuracy per se. There is thus available, in the process testing literature, a wealth of techniques and formats from which the classroom teacher, textbook author, or others should be in a position to choose both diagnostically-oriented and more generalized assessment techniques to best suit the process measurement purposes at issue at specified points in the teaching program.

From a practical standpoint, the curriculum-free nature of proficiency testing has the very great advantage of making it possible to develop single instruments or batteries of instruments that have wide-ranging applicability across a variety of instructional levels, textbook series, or other aspects of the instructional program. Process measurement, on the other hand, is appreciably more restricted in that it is focused, by definition, on determining the student's acquisition or lack of acquisition of various aspects of course

content to which the student has been exposed through textbook study, classroom sessions, or other instructional means. To the extent that present-day instructional programs vary among themselves in content, sequencing, and other characteristics, it is not possible to specify or develop uniform achievement (process) tests for use in these programs, even though this is a reasonable and practical option in countries having a standardized national language syllabus.

In the U.S. setting at the present time, there would appear to be three possible levels at which useful steps could be taken to develop and improve process testing practices: (1) within individual classrooms (i.e., at the teacher or local school district level); (2) within publishers' textbook programs; and (3) (on an advisory if not mandatory basis) at the national level through initiatives by language teachers' organizations or other appropriate entities.

In the shorter version of this paper, it was suggested to be appropriate for the NIE to consider providing developmental assistance in the first two of these areas. At the individual teacher/school district level, one suggestion was for the development and dissemination of a simple, highly specific "how-to" manual on the preparation and use of classroom achievement tests, including the presentation of a large number of question-type "frames" or patterns that the teacher could easily follow in preparing suitable tests on a local basis.

At the publisher level, two possible areas for NIE consideration were identified. The first of these was a project to develop and publicize--within the language teaching profession itself and with the assistance of measurement resource persons as necessary--a detailed "code of good measurement practice" for textbook-related test materials, which state supervisors, local school districts, or others could use as an important reference in connection with textbook review and selection. (Publishers losing one or more statewide adoptions

as a consequence of identified deficiencies in the accompanying testing materials would have a real and immediate incentive to improve the quality of this aspect of their textbook programs.)

A second suggestion (not incompatible with the first) was to sponsor one or more intensive and well-publicized colloquia on the nature, role, and potential instructional benefits of systematic, textbook-based achievement testing, in which publishers as well as foreign/second language representatives and testing specialists would be asked to participate. It was suggested that either or both of these activities might be arranged and coordinated through the American Council on the Teaching of Foreign Languages (ACTFL), an approach which would provide both professional stature and across-languages emphasis to these undertakings.

Although activity in each of the preceding areas would still be considered of value, a recent initiative at the national level has been made by ACTFL in connection with a "realistic learning outcomes" project, for which external support is being sought. The nature of the work proposed by ACTFL is such as to have substantial potential ramifications for proficiency-oriented process measurement and would be considered to take priority over (and ultimately appreciably influence the nature of) development work at the classroom and publisher levels.

Briefly, the "realistic learning outcomes" project involves literature review, expert consultation, and other activities aimed at developing three levels of description of language learning goals/objectives: (1) general proficiency statements not related to a particular language or course of study; (2) somewhat more detailed learning outcomes that would be language-specific and would relate to one or more identified curricula or textbook

series; and (3) instructional behavioral objectives, also language-specific and even more highly detailed, that would serve at the "individual item" level for the corresponding process tests and that could be used flexibly by the teachers (or textbook publishers) in setting up and assessing the particular instructional sequences they intend to follow.

The proposed ACTFL project would be considered to merit attention both in its own right and as a possible superstructure from which a considerable amount of useful additional development and dissemination work could flow, including both textbook publisher initiatives to incorporate the learning outcomes identified in the project and "how-to" test development manuals for the classroom teacher, based on the same objectives. It should also be noted that the "realistic learning outcomes" project is in keeping with an additional recommendation of the President's Commission urging the "[establishment of] language proficiency achievement goals for the end of each year of study at all levels, with specific attention to speaking proficiency."

SUMMARY AND ADDITIONAL RECOMMENDATIONS

The suggestions for potential NIE activity in foreign/second language testing research and development that have been discussed in the preceding pages may be summarized as follows:

(1) In the area of input (prognostic) assessment, some consideration should be given to supporting research studies aimed at further elucidating the nature, role, and influence of "learner variables" on language acquisition, and including the concomitant development of background measures, attitude/motivation scales, and other types of instruments capable of measuring these variables in a valid, consistent, and, to the extent possible, easily administerable way. Full realization of the practical benefits available from this type of assessment

would necessitate instructional flexibilities within the teaching programs themselves that are not generally available at the present time.

(2) Of the three measurement areas discussed, the development and utilization of new and much more highly "real-life performance" referenced proficiency tests is considered of the greatest priority, both as measures of the intended proficiency-based language learning goals and as instructional benchmarks toward which language curricula can and should be focused. A sequential program of test development, trial administration and validation, and subsequent ongoing accumulation and reporting of interpretive information obtained from the field is suggested as a psychometrically appropriate and workable approach to such development, desirably through the administrative framework of an institution-based "research and development center" which would provide for the necessary conceptual and practical continuity across all project phases.

(3) With respect to course-of-instruction (process) measurement, possible areas of useful development activity include the preparation of detailed training materials for teachers in the preparation of proficiency-oriented classroom tests, as well as the sponsoring of teacher/tester/publisher colloquia concerning exemplary achievement measurement practices. An important framework for these and other related activities could be provided by the "realistic learning outcomes" project proposed by ACTFL. This project would involve the detailed working out and dissemination of proficiency-based learning objectives at a level of specificity that would be of considerable practical utility in instructional planning and in the development of the corresponding process tests at both individual-teacher and textbook program levels.

Several other areas may briefly be identified as meriting consideration

for NIE support, either as worthwhile undertakings in their own right or as activities usefully related to one or more of the projects previously discussed.

(1) There is a need for both language researchers and test developers to have available, for use in a variety of studies and other projects, a series of standardized data gathering instruments for obtaining accurate and uniform biographical and language-learning history information from the study participants (e.g., age, sex, native language, language background of parents, history of formal language study, informal language contacts, etc.). At present, there is little or no consistency across reported studies in the way in which this information is obtained and represented, a situation that not only reflects considerable inefficiency in the research process (with each individual researcher "reinventing the wheel" in developing the necessary instruments) but also--and more importantly--makes it difficult or impossible to properly evaluate the research results on an across-study basis.

A much preferable approach would be to convene a moderately sized (6-7 member) working group of language researchers and other relevant persons to develop a uniform, general-purpose biographical/language history questionnaire which, following trial administration and revision as necessary, could be used as a standard data gathering instrument in a wide number of research studies and other language-related projects. A number of studies under way or being planned as components of the "Language Skills Attrition Project" at the University of Pennsylvania under the general supervision of Richard Lambert and G. Richard Tucker (Lambert and Freed, 1980) depend in large part on the obtaining of detailed and consistent biographical/language background information, and members of this project group would be in an ideal position to plan, develop, and disseminate the necessary instruments as one aspect of the attrition project work.

(2) The nature and value of self- and second-party language proficiency appraisal instruments has already been described in connection with their use as construct validation criteria for proficiency test development. Both types of instruments are of considerable potential usefulness in their own right as indices of language proficiency in situations where it is impossible or highly impractical to administer actual proficiency tests. The "Global Understanding" study carried out at ETS as part of the Council on Learning's Education and the World View project (Barrows, Klein, and Clark, in press) used proficiency self-assessment scales for second-language speaking, listening comprehension, and reading in its nationwide survey of college students. These scales (following initial validation against measured language performance) were found extremely useful as "stand-ins" for the direct proficiency tests, and provided information on the general level of language proficiency of the survey groups that could not otherwise have been obtained.

Further development of the self-appraisal scales, revalidation on suitable examinee groups, and dissemination for in-field research and other uses, would be considered a very worthwhile activity, either as part of the larger test development program previously described or as a separate project.

(3) In view of the large amount of highly relevant curriculum development and proficiency assessment work that is being carried out through the Council of Europe, the support of travel and related costs to facilitate interaction between Eurpoean and American language testing researchers and developers would be recommended, either through American visits to institutions abroad or lecture tours by invited visitors to the U.S. A very useful related undertaking would be to continue data gathering activities of the type carried out by Randall Jones (1980), who conducted an international survey study of language testing

research projects resulting in some 286 entries (with brief descriptions of the projects) as of the time of compilation. Periodic updating of this type of survey information would be of appreciable interest and value to language testing researchers in the United States.

References

Barrows, Thomas, Stephen Klein, and John L. D. Clark. What Students Know and Believe About Their World (in press).

Brod, Richard I., ed. Language Study for the 1980's: Reports of the MLA-ACLS Language Task Forces. New York: Modern Language Association, 1980.

Bryan, Miriam M. "The MLA-Cooperative Foreign Language Tests: Tests with a New Look and a New Purpose," DFL Bulletin, vol. 6, no. 2, Dec. 1966, pp. 6-8.

Canale, Michael and Merrill Swain. Communicative Approaches to Second Language Teaching and Testing. Toronto: Ontario Ministry of Education, 1979.

Carroll, Brendan J. Testing Communicative Performance. Oxford: Pergamon Press, 1980.

Carroll, John B. "The Prediction of Success in Intensive Foreign Language Training." In Robert Glaser, ed., Training Research and Education. Pittsburgh: University of Pittsburgh Press, 1962, pp. 87-136.

Carroll, John B. and Stanley M. Sapon. Modern Language Aptitude Test. New York: The Psychological Corporation, 1959.

Carroll, John B. and Stanley M. Sapon. Modern Language Aptitude Test--Elementary. New York: The Psychological Corporation, 1967.

Clark, John L. D. Foreign Language Testing: Theory and Practice. Philadelphia: Chilton/Center for Curriculum Development, 1972.

References (cont.)

Clark, John L. D. "MLA-Cooperative Foreign Language Tests," Journal of Educational Measurement, vol. 2, no. 2, Dec. 1965, pp. 234-244.

Clark, John L. D. "Toward a Common Measure of Speaking Proficiency." In James R. Frith, ed., Measuring Spoken Language Proficiency. Washington, D.C.: Georgetown University Press, 1980.

Clark, John L. D. and Spencer S. Swinton. The Test of Spoken English as a Measure of Communicative Ability in English-Medium Instructional Settings. TOEFL Research Report No. 7. Princeton, N.J.: Educational Testing Service, 1980.

Cohen, Andrew D. Testing Language Ability in the Classroom. Rowley, Mass.: Newbury House, 1980.

Gardner, Robert C., L. Gliksman, and P.C. Smythe. "Attitudes and Behavior in Second Language Acquisition: A Social Psychological Interpretation," Canadian Psychological Review, vol. 19, 1978, pp. 173-186.

Gardner, Robert C. and Wallace E. Lambert. Attitudes and Motivation in Second Language Learning. Rowley Mass.: Newbury House, 1972.

Harris, David P. Testing English as a Second Language. New York: McGraw-Hill, 1969.

Jones, Randall L. An International Survey of Research in Language Testing: 1977-79. Provo, Utah: Brigham Young University, 1980.

Lado, Robert. Language Testing: The Construction and Use of Foreign Language Tests. London: Longmans, 1961.

References (cont.)

Lado, Robert. "Scope and Limitations of Interview-Based Language Testing: Are We Asking Too Much of the Interview?" In John L. D. Clark, ed., Direct Testing of Speaking Proficiency: Theory and Application. Princeton, N.J.: Educational Testing Service, 1978.

Lambert, Richard D. and Barbara Freed, eds. Proceedings of the Conference on the Attrition of Language Skills (in press [Newbury House]).

Oller, John W., Jr. Language Tests at School. London: Longman, 1979.

Pimsleur, Paul. The Pimsleur Language Aptitude Battery. New York: Harcourt, Brace, 1966.

Pimsleur, Paul. Pimsleur Modern Foreign Language Proficiency Tests. New York: Harcourt, Brace, 1967.

President's Commission on Foreign Language and International Studies. Strength Through Wisdom: A Critique of U.S. Capability. Washington, D.C.: U.S. Government Printing Office, 1979.

Rice, Frank A. "The Foreign Service Institute Tests Language Proficiency," Linguistic Reporter, vol. 1, no. 2, 1959, p. 4.

Sollenberger, Howard E. "Development and Current Use of the FSI Oral Interview Test." In John L. D. Clark, ed., Direct Testing of Speaking Proficiency: Theory and Application. Princeton, N.J.: Educational Testing Service, 1978, pp. 1-12.

References (cont.)

Starr, Wilmarth H. "MLA Foreign Language Proficiency Tests for Teachers and Advanced Students," PMLA, vol. 77, no. 4, part 2, 1962, pp. 1-12.

Trim, J.L.M. "Draft Outline of a European Unit/Credit System." Systems Development in Adult Language Learning. Strasbourg: Council of Europe, 1973.

Trim, J.L.M. "Towards a Situational Definition of Language Contents: A Schematic Inventory of Language Roles, Activities, and Uses." Modern Language Learning in Adult Education. CCC/EES (71) 135. Strasbourg: Council of Europe, 1971.

Valette, Rebecca M. Modern Language Testing: A Handbook. New York: Harcourt, Brace, 1967.

Valette, Rebecca M. Modern Language Testing: A Handbook (Second Edition). New York: Harcourt, Brace, 1977.

Van Ek, J.A. Systems Development in Adult Language Learning: The Threshold Level. Strasbourg: Council of Europe, 1975.

Wilds, Claudia P. "The Oral Interview Test." In Randall L. Jones and Bernard Spolsky, eds., Testing Language Proficiency. Arlington, Va. [now Washington, D.C.]: Center for Applied Linguistics, 1975, pp. 29-44.